Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

Research Article

Cancer Immunology Research

Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types

Je-Keun Rhee[1,2], Yu Chae Jung[3], Kyu Ryung Kim[1,2], Jinseon Yoo[1,2], Jeeyoon Kim[1,2], Yong-Jae Lee[1,2], Yoon Ho Ko[4], Han Hong Lee[5], Byoung Chul Cho[6], and Tae-Min Kim[1,2]

Abstract

Surgical archives of tumor specimens are often impure. The presence of RNA transcripts from nontumor cells, such as immune and stromal cells, can impede analyses of cancer expression profiles. To systematically analyze the impact of tumor purity, the gene expression profiles and tumor purities were obtained for 7,794 tumor specimens across 21 tumor types (available in The Cancer Genome Atlas consortium). First, we observed that genes with roles in immunity and oxidative phosphorylation were significantly inversely correlated and correlated with the tumor purity, respectively. The expression of genes implicated in immunotherapy and specific immune cell genes, along with the abundance of immune cell infiltrates, was substantially inversely correlated with tumor purity. This relationship may explain the correlation between immune gene expression and mutation burden, highlighting the need to account for tumor purity in the evaluation of expression markers obtained from bulk tumor transcriptome data. Second, examination of cluster membership of gene pairs, with or without controlling for tumor purity, revealed that tumor purity may have a substantial impact on gene clustering across tumor types. Third, feature genes for molecular taxonomy were analyzed for correlation with tumor purity, and for some tumor types, feature genes representing the mesenchymal and classical subtypes were inversely correlated and correlated with tumor purity, respectively. Our findings indicate that tumor purity is an important confounder in evaluating the correlation between gene expression and clinicopathologic features such as mutation burden, as well as gene clustering and molecular taxonomy. *Cancer Immunol Res; 6(1); 87–97. ©2017 AACR.*

Introduction

Solid tumor tissues are comprised of cellular components originating from various types of cancerous and noncancerous tissues, the latter including immune, stromal, endothelial, and epithelial cells (1). Such noncancerous cells are regarded as common contaminants in tumor admixtures and constitute a substantial fraction of tumor masses. They are also important in carcinogenesis. Stromal or mesenchymal cells may enhance tumor growth and influence the response of cancers (2), whereas immune cells, such as tumor infiltrating cytotoxic T lymphocytes, may inhibit tumor growth (3, 4). The extent of nontumor cell contamination or tumor purity (the proportion of tumor cells in the mixture) can be estimated by histologic examination or computational methods using various genomic resources (5, 6). Standard tumor sampling using surgical specimens may achieve a tumor purity that exceeds 70%, but it can often be lower, leading to systematic biases in tumor genome analyses (5). This contamination can complicate tumor transcriptome analyses due to the cancer masses representing a mixture of RNA transcripts originating from tumor cells and noncancerous cells. On the other hand, reports state that the impurity of tumors can be leveraged for immunologic insights (7, 8). In these studies, the relative proportion of tumor-infiltrating lymphocytes and other immune cells can be estimated from the tumor transcriptome by deconvolution methods.

Cancer transcriptome analyses, especially those focused on immune-related genes, must take into account the tumor purity because immune cells are a major fraction of noncancerous cells in the tumor mass. However, the impact of tumor purity on the expression of immune-related genes remains unclear. Immune checkpoint blockade using monoclonal antibodies against cytotoxic T lymphocyte–associated protein 4 (CTLA-4; ref. 9) and programmed cell death protein 1/programmed death-ligand 1 (PD-1/PD-L1; ref. 10) produces durable clinical responses for a number of solid tumors including melanoma (11) and non–small cell lung cancer (12). The remarkable success rate of clinical

[1]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Korea. [2]Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul, Korea. [3]Department of IT Engineering, Sookmyung Women's University, Seoul, Korea. [4]Division of Medical Oncology, Department of Internal Medicine, College of Medicine, The Catholic University of Korea, Seoul, Korea. [5]Division of Gastrointestinal Surgery, Department of Surgery, College of Medicine, The Catholic University of Korea, Seoul, Korea. [6]Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Korea.

Note: Supplementary data for this article are available at Cancer Immunology Research Online (http://cancerimmunolres.aacrjournals.org/).

J.-K. Rhee and Y. C. Jung share first authorship of this article.

Corresponding Authors: Tae-Min Kim, The Catholic University of Korea. 505, Banpo-Dong, Seocho-Ku, Seoul 137-701, Korea. Phone: 82-2-2258-7352; Fax: 82-2-537-0572; E-mail: tmkim@catholic.ac.kr; Byoung Chul Cho, Department of Internal Medicine, Yonsei Cancer Center, Division of Medical Oncology, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu 120-752, Seoul, Korea. E-mail: cbc1971@yuhs.ac

doi: 10.1158/2326-6066.CIR-17-0201

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

trials utilizing this approach is tempered by the reality that only a fraction of patients will respond to the treatment. Thus, a need exists to identify biomarkers that can select patients with a better chance of achieving clinical benefits with cancer immunotherapy. The mutation burden of tumor genomes is considered a reliable predictor for the response to immunotherapy (9, 10, 13). Along with the somatic mutations that may produce neoepitopes eliciting antitumor T cell–mediated responses (14), the expression of immune genes for the targets or proxies of immunotherapy (e.g., CTLA-4 and PD-1/PD-L1) has been proposed (15–18). Other expression-based markers representing the activity of tumor-infiltrating T lymphocytes, such as granzyme A (encoded by *GZMA*), perforin-1 (encoded by *PRF1*; ref. 19), and CD8α (encoded by *CD8A*; ref. 20), may also be potentially useful as predictive markers for immunotherapy. The expression of such immune-related genes originating from noncancerous immune cells can be associated with the tumor purity. The differential expression between tumor and nontumor tissues may be biased, as demonstrated for immune-related genes such as those encoding CTLA-4 and its ligand, CD86 (5).

An early study on tumor purity has reported the potential influence of the tumor purity on the coexpression network, clustering-based tumor subtyping or molecular taxonomy, and the identification of differentially expressed genes (5). This study demonstrated that the coexpressed gene pairs often arise merely due to the tumor purity, highlighting the need for purity adjustment. However, it is largely unknown which gene pairs or tumor types are subject to purity adjustment and how to estimate the impact of tumor purity on gene clustering. Aran and colleagues showed that the certain tumor subtypes, such as mesenchymal subtypes of malignant brain tumors, were characterized by low purity, and the correlation of the feature gene expression with tumor purity showed a bimodal distribution (5). Similar reports have been published suggesting that the expression-based molecular taxonomy and risk assessment can be biased by tumor purity (21), and the expression signals representing mesenchymal tumor subtypes have been attributed mainly to stromal cells instead of tumor cells (22, 23).

In this study, we scrutinized the large PanCancer transcriptome database containing 7,794 tumor specimens and 21 tumor types from The Cancer Genome Atlas (TCGA) consortium. Tumor purity for the corresponding tumor samples was obtained in a PanCancer tumor purity study (5). Three aspects of cancer transcriptome analyses were investigated for the potential impact of tumor purity—the correlation of tumor purity with individual genes (especially those related to immunity), gene clustering, and molecular taxonomy. Although purity has been previously evaluated as a potential confounding factor in gene expression analyses (5), we further investigated the impact of tumor purity adjustment on the correlation between the expression of genes, including known immune markers and mutation burden. We also evaluated the impact of tumor purity on gene clustering and observed that a substantial number of gene pairs lost their cluster membership concordance after purity adjustment, and the extent was variable across tumor types. A substantial inverse correlation or correlation with tumor purity was observed for feature genes, such as those representing mesenchymal or classic tumor subtypes, suggesting that tumor purity also represents a major confounding factor in expression-based molecular taxonomy.

Methods and Materials

Datasets

We obtained RNA sequencing (RNA-seq)–based gene expression profiles of 7,794 tumor specimens across 21 tumor types from the TCGA pan-cancer consortium (https://cancergenome.nih.gov/). The tumor types examined included 79 adenocortical carcinoma (ACC), 408 bladder urothelial carcinoma (BLCA), 1,100 breast invasive carcinoma (BRCA), 306 cervical and endocervical carcinoma (CESC), 287 colon adenocarcinoma (COAD), 166 glioblastoma multiforme (GBM), 522 head and neck squamous cell carcinoma (HNSC), 66 kidney chromophobe (KICH), 534 kidney renal clear cell carcinoma (KIRC), 291 kidney renal papillary cell carcinoma (KIRP), 530 lower grade glioma (LGG), 373 liver hepatocellular carcinoma (LIHC), 517 lung adenocarcinoma (LUAD), 501 lung squamous cell carcinoma (LUSC), 307 ovarian serous cystadenocarcinoma (OV), 498 pancreatic adenocarcinoma (PRAD), 95 rectum adenocarcinoma (READ), 471 skin cutaneous melanom (SKCM), 509 thyroid carcinoma (THCA), 177 uterine corpus endometrial carcinoma (UCEC), and 57 uterine carcinosarcoma (UCS). We also obtained tumor purity for these selected cases as the median of four types of tumor purity estimated from different histologic or genomic resources, as previously used as a consensus purity measure (5).

Resource for immune cell infiltrates and mutation burdens

Marker genes representing 20 immune cell types were obtained from a list of commercial gene panel (nCounter Human PanCancer Immune Profiling Panel; NanoString Technologies). Per immune cell type, three to eight representative, immune-related genes were selected from the panel gene list and their expression in TCGA specimens were investigated for correlation with tumor purity. We also obtained the relative abundance of immune cell infiltrates for individual TCGA tumor specimens as estimated by two deconvolution algorithms, TIMER (7) and CIBERSORT (8). To correlate gene expression with the burdens of somatic mutation and predicted neoantigens, we used the expression profiles of 3,588 tumor specimens, for which the number of neoantigens has been predicted (19). For somatic mutations, we calculated the natural or unadjusted correlation between the expression of all the genes examined and the mutation abundance, as well as the partial correlation controlling for the tumor purity. We then focused on known immune genes *GZMA*, *PRF1*, *PDCD1* (encodes PD-1), *CD274* (encodes PD-L1), *CTLA4*, and *CD8A*, as well as genes related with mutation abundance such as *APOBEC3B* and *MLH1*. In case of *MLH1*, the correlation with tumor purity was calculated across three tumor types with frequent microsatellite instability (COAD, STAD, and UCEC). CYT (cytolytic activity) was calculated as the geometric mean of *GZMA* and *PRF1* expression, as previously indicated (19). All statistical analyses were done using R software (https://www.r-project.org/) unless specified otherwise.

Gene set enrichment analysis

To identify the molecular functions enriched in the tumor purity-correlated and -inversely correlated genes, we used a pre-ranked version of a gene set enrichment analysis (GSEA; ref. 24). For all the genes examined, Spearman correlations were calculated between their expression and tumor purity to generate a ranked list of genes. The pre-ranked version of GSEA was performed using the Gene Ontology terms (MSigDB, c5 category; http://software.

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

broadinstitute.org/gsea/msigdb) or c2cp category (curated pathways) as functional gene sets. GSEA was also used to measure the extent of enrichment for feature genes toward correlation or inverse correlation with tumor purity.

Survival analysis

Per tumor type, Kaplan–Meier survival curves were generated by partitioning the patients into high and low CYT gene expression using the median. Log-rank test was used to estimate the significance of survival between patients with high and low CYT expression. To control for the tumor purity, Kaplan–Meier survival curves were also generated using the residual of CYT expression after linear regression against the tumor purity. The residuals were also used to evaluate the association between purity-adjusted CYT and survival by log-rank tests. $P < 0.05$ was significant.

Purity-adjusted gene clustering

For each tumor type, we first identified a subset of genes with variable expression (i.e., 5,000 genes with top median absolute deviation) and assigned them into six gene clusters by consensus clustering (25). The number of gene clusters for consensus clustering ($k = 6$) was determined by the manual examination of consensus cumulative distribution function plots across the tumor types. Using the results of purity-unadjusted gene clustering, we discriminated the possible gene pairs into CM and CMM pairs (cluster-matched and cluster-mismatched, respectively) according to cluster membership of genes in the pairs. For purity-adjusted gene clustering, a gene-wise distance matrix was generated using the value of 1 − correlation (partial Spearman correlation between the expression of two genes in a given pair controlled for the tumor purity) as a distance measure. The ppcor R package (26) was used to calculate partial correlation by controlling for the tumor purity. Using the results of purity-adjusted gene clustering, CM and CMM gene pairs were further discriminated into CM-C/CMM-C (C for "concordant") and CM-D/CMM-D (D for "discordant"). For example, if two genes in a pair belong to the same cluster, the corresponding gene pair is CM. The CM gene pairs were further annotated as CM-C if the two genes in the pair fell into the same cluster after purity adjustment or, otherwise annotated as CM-D. CMM-C and CMM-D gene pairs were annotated similarly. We obtained 39,240 protein–protein interaction (PPI) gene pairs from public database of HPRD (Human Protein Reference Database; ref. 27). The relative abundance of PPI gene pairs was calculated across the four categories of gene pairs.

Selection of feature genes

For GBM, we obtained 840 feature genes that were used to classify the TCGA GBM cases into four subtypes (proneural, neural, classic, and mesenchymal; ref. 28). Four subtype-specific GBM feature genes were measured for their correlation with tumor purity using pre-ranked GSEA (24). To demonstrate the impact of tumor purity on GBM molecular taxonomy, the feature genes were partitioned into 400 purity-correlated and 400 purity-inversely correlated genes. The 173 GBM cases with annotated subtypes were subject to hierarchical clustering using these two subsets of feature genes. Hierarchical clustering was done using 1 − Pearson correlation as distance with average linkage. We further obtained the tumor subtypes for additional five tumor types (BRCA, COAD, HNSC, LUAD, and LUSC). To identify the feature genes in each tumor type, we used ClaNC R package (29). The number of feature genes was determined as the smallest number of genes with the lowest cross-validation and prediction error as reported by ClaNC. The extent of correlation with tumor purity was evaluated for individual sets of feature genes per tumor type by pre-ranked GSEA.

Results

Relationship between tumor purity and gene expression

We first examined the correlation between gene expression and tumor purity (Fig. 1A). Spearman correlation (ρ) was calculated for the expression of individual genes and tumor purity across 7,794 tumors encompassing 21 tumor types. A tendency of negatively skewed distribution of the correlation coefficients suggested that the expression of many genes was inversely correlated with tumor purity. GSEA was done to identify molecular functions enriched with genes that were correlated or inversely correlated with tumor purity (Table 1). Mitochondria- and immune-related molecular terms (MSigDB c5 Gene Ontology or GO categories) were correlated and inversely correlated with tumor purity, respectively. Similar molecular terms such as "respiratory electron transport" and "cytokine-cytokine receptor interaction" were identified using other functional gene sets (MSigDB c2cp as Curated Pathways; Supplementary Table S1). The correlation of the genes belonging to the top 10 correlated or inversely correlated molecular functions are shown in Fig. 1B ($n = 175$ and 1,140 genes as aggregates of genes in 10 mitochondria- and 10 immune-related GO categories in Table 1 as red and blue, respectively) and suggests that immune-related genes comprise a major fraction of genes at the skewed edge of the distribution of the genes that were inversely correlated with tumor purity. We counted the number of occurrences in 10 immune-related GO categories listed in Table 1 for the 1,140 immune-related genes in Fig. 1B. The genes with a higher number of calls in the 10 immune-related GO categories were inversely correlated to a greater extent with the tumor purity than those with a lower number of calls (Supplementary Fig. S1).

Because the tumor purity used in this study was mainly the consensus of purity estimates from multiple genomic and pathologic resources (5), the same analyses were done on four types of tumor purities from different genomic resources or histology (copy number, gene expression, DNA methylation, and pathologic examination). We consistently observed enrichment of immune-related genes toward the inverse correlation with tumor purity, regardless of the genomic resource to estimate that tumor purity (Supplementary Fig. S2). We also examined the relationship between gene expression and tumor purity for individual tumor types. The distribution of correlation coefficients was similar across the 21 tumor types (Supplementary Fig. S3). Although the immune-related genes were consistently enriched in genes inversely correlated with tumor purity across the 21 tumor types (Supplementary Table S2), the genes correlated with tumor purity across tumor types were enriched for diverse molecular functions: oxidative phosphorylation/mitochondria-related genes for BRCA, KIRC, LIHC, PRAD, and THCA tumors; ribosome/translation-related genes for COAD, HNSC, KICH, LUAD, READ, UCEC, SKCM, UCEC, and UCS tumors; chromosome/DNA replication-related genes for ACC, GBM, and LUSC tumors; histone methylation-related genes for BRCA tumors; cilium/

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

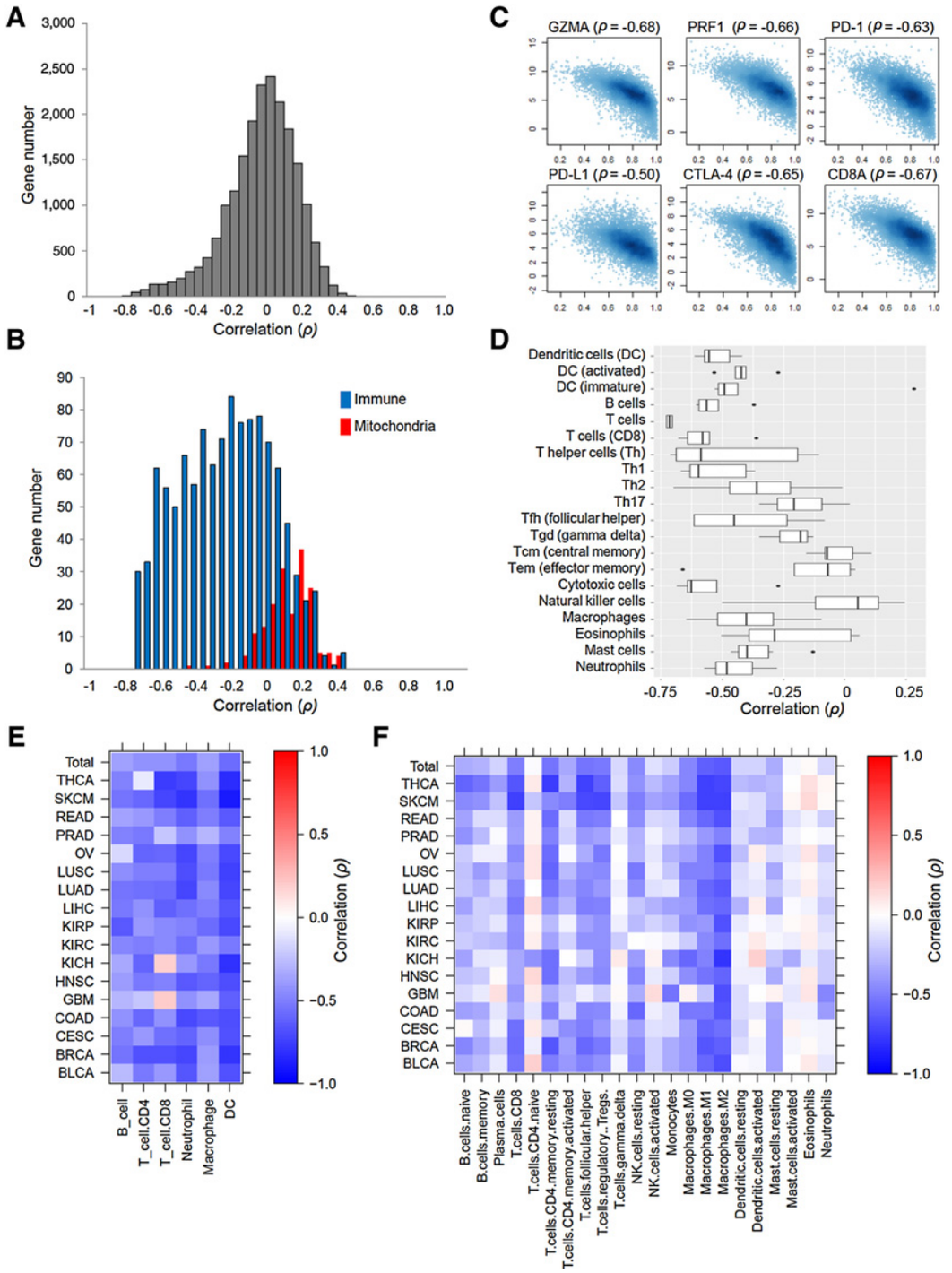

Figure 1.

The inverse correlation of immune-related genes with tumor purity. **A,** The distribution of correlation coefficients (*ρ*, Spearman correlation) between the gene expression and tumor purity is shown. **B,** The correlation coefficients is shown in a histogram for the selected immune- and mitochondria-related genes. Immune: 1,140 genes, red bars; mitochondria: 175 genes, blue bars. **C,** Scatter plots show the expression of genes and tumor purity (*y* and *x*-axes, respectively) for six immune marker genes selected. The expression levels are shown in $\log_2$ scale. **D,** Three to eight genes representing 20 immune cell types were selected, and correlation with tumor purity is shown. **E,** The correlation with tumor purity is shown for the abundance of six immune cell infiltrates as estimated by TIMER algorithm in a heatmap. Top row (total): the correlation measured across the total dataset and the correlation for individual tumor types are shown below. A color indicator shows the level of correlation with tumor purity. **F,** Correlation with tumor purity for 22 types of immune infiltrates whose abundance was measured by CIBERSORT algorithm. The annotation of immune cell infiltrates is according to the output of the used algorithms.

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

Table 1. Molecular functions correlated or inversely correlated with tumor purity

Correlation	Functions (MSigDB c5, GO categories)	Genes	ES	NES	*P* value	FDR	FWER
Correlated	GO_MITOCHONDRIAL_RESPIRATORY_CHAIN_COMPLEX_ASSEMBLY	74	0.7	2.9	0	0	0
	GO_MITOCHONDRIAL_RESPIRATORY_CHAIN_COMPLEX_I_ASSEMBLY	55	0.6	2.8	0	0	0
	GO_NADH_DEHYDROGENASE_COMPLEX_ASSEMBLY	55	0.6	2.8	0	0	0
	GO_NADH_DEHYDROGENASE_COMPLEX	42	0.7	2.8	0	0	0
	GO_MITOCHONDRIAL_RESPIRATORY_CHAIN_COMPLEX_I_BIOGENESIS	55	0.6	2.8	0	0	0
	GO_MITOCHONDRIAL_ELECTRON_TRANSPORT_NADH_TO_UBIQUINONE	41	0.7	2.7	0	0	0
	GO_INNER_MITOCHONDRIAL_MEMBRANE_PROTEIN_COMPLEX	101	0.6	2.7	0	0	0
	GO_RESPIRATORY_CHAIN	78	0.6	2.6	0	0	0
	GO_NADH_DEHYDROGENASE_ACTIVITY	37	0.7	2.6	0	0	0
	GO_MITOCHONDRIAL_PROTEIN_COMPLEX	131	0.5	2.6	0	0	0
Inversely correlated	GO_ADAPTIVE_IMMUNE_RESPONSE	252	−0.8	−3.1	0	0	0
	GO_CELLULAR_RESPONSE_TO_INTERFERON_GAMMA	119	−0.8	−3.0	0	0	0
	GO_RESPONSE_TO_INTERFERON_GAMMA	141	−0.8	−3.0	0	0	0
	GO_REGULATION_OF_INTERFERON_GAMMA_PRODUCTION	93	−0.8	−3.0	0	0	0
	GO_POSITIVE_REGULATION_OF_CELL_ACTIVATION	284	−0.7	−2.9	0	0	0
	GO_REGULATION_OF_LEUKOCYTE_PROLIFERATION	202	−0.7	−2.9	0	0	0
	GO_REGULATION_OF_T_CELL_PROLIFERATION	144	−0.8	−2.9	0	0	0
	GO_INFLAMMATORY_RESPONSE	445	−0.7	−2.9	0	0	0
	GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	365	−0.7	−2.9	0	0	0
	GO_REGULATION_OF_LEUKOCYTE_MEDIATED_IMMUNITY	155	−0.8	−2.9	0	0	0

NOTE: ES, NES, FDR, and FWER are enrichment score, normalized ES, false discovery rate, and family-wise error rate, respectively.

motility-related genes for CESC, KIRP, and OV tumors; and glutamate receptor-related genes for LGG tumors.

Next, we selected six immune markers, including genes encoding cytotoxic T lymphocyte differentiation markers (granzyme A, perforin, and CD8α) and genes encoding multiple immunosuppressive checkpoints (PD-1, PD-L1, and CTLA-4) as targets of immunotherapy (4, 19). The expression of these genes was inversely correlated with tumor purity ($\rho < -0.5$; $P < 2.2 \times 10^{-16}$; Fig. 1C). For correlations of specific immune cells, we further selected known marker genes for 20 immune cell types. Figure 1D shows the distribution of correlations with the tumor purity for three to eight marker genes per immune cell type (the list of marker genes is available in Supplementary Table S3). Most of the marker genes for particular immune cell types were inversely correlated with tumor purity ($\rho < 0$ for 89 genes out of 100 immune cell marker genes selected). Those representing dendritic cells, B cells, T cells, and cytotoxic cells were the most inversely correlated (median of $\rho < -0.5$). The inverse correlation for the genes representing memory T cells and natural killer cells was observed to a lesser extent. Because investigation based on a small number of genes may be subject to a selection bias, we further investigated the relative abundance of immune cell filtrates measured by expression-based deconvolution for their correlation with tumor purity. Figure 1E shows a heatmap representing the correlations between the abundance of six immune cell infiltrates (B cells, $CD4^+$ and $CD8^+$ T cells, neutrophils, macrophages, and dendritic cells, as estimated by TIMER algorithm; ref. 7) and tumor purity. Overall, inverse correlations with tumor purity were observed for the majority of immune cell infiltrates across tumor types (98.1% of correlations were inverse correlations, $\rho < 0$). This was also true for the abundance of 22 types of immune cells, as estimated by CIBERSORT algorithm (8), where 88.3% of correlations were inverse correlations (Fig. 1F). Substantial inverse correlation was observed for $CD8^+$ T cells and M2 macrophages. These results indicate that the immune-related molecular functions represent the major functional category whose member genes are inversely correlated with tumor purity, which is also true for known immune marker genes and the relative abundance of immune cell infiltrates.

Immune marker gene expression and mutation burden

To examine the impact of tumor purity on the expression of immune genes, we focused on the previously reported correlation between the immune marker gene expression and mutation burden (19). Consistent with the previous report, an expression-based measure of cytotoxic T lymphocytes infiltrates, CYT (cytolytic activity, a geometric mean of *GZMA* and *PRF1* expression), was significantly correlated with the mutation burden ($\rho = 0.19$, $P < 2.2 \times 10^{-16}$; Fig. 2A). However, these two parameters (CYT and mutation abundance) were also significantly inversely correlated with tumor purity ($\rho = -0.70$ and -0.20, respectively; both $P < 2.2 \times 10^{-16}$; Fig. 2B and C), suggesting that the correlation between CYT and mutation burdens may have come from their relationship with the tumor purity. The abundance of predicted neoantigens was significantly correlated with the mutation abundance ($\rho = 0.90$, Supplementary Fig. S4A). As expected, substantial inverse correlation and correlation were observed between the number of neoantigens with tumor purity ($\rho = -0.21$, Supplementary Fig. S4B) and with CYT ($\rho = 0.19$, Supplementary Fig. S4C), respectively. These findings suggest that the previously reported correlation between CYT and the burdens of somatic mutation or neoantigens (19) may likely result from the correlation of these features with tumor purity.

To further investigate the impact of tumor purity on the observed correlation between gene expression and mutation burdens, we applied partial correlations to control for tumor purity (5). For the comparison, we calculated two types of correlation—the expression of all the genes and mutation burdens with or without controlling for tumor purity (adjusted and unadjusted/natural correlation, respectively). A scatter plot of the distribution of the two types of correlations for all genes (purity-adjusted and -unadjusted/natural correlation for *y*- and *x*-axes, respectively; Fig. 2D) revealed that these two correlations were

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

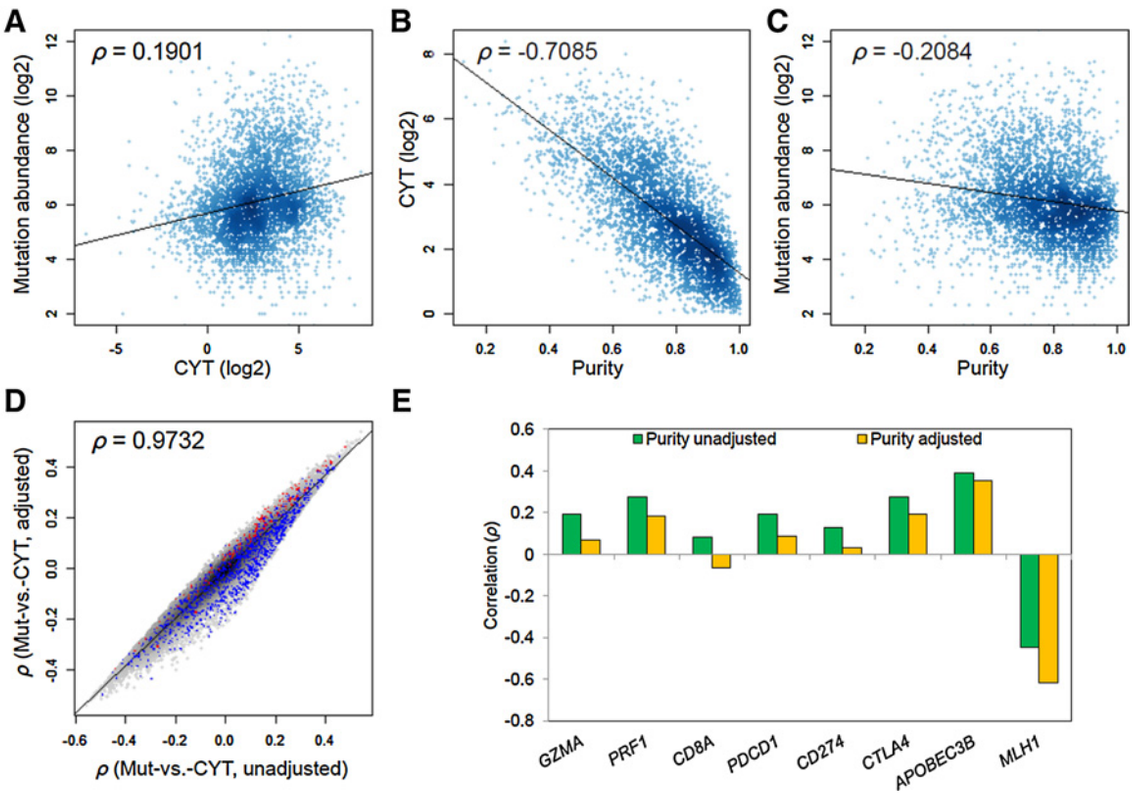

Figure 2.

Relationship between the expression of immune-related genes and burdens of somatic mutation. **A,** Scatter plot showing CYT (a geometric mean of *GZMA* and *PRF1* expression; *x*-axis in $\log_2$ scale) and the number of somatic mutations (*y*-axis; $\log_2$ scale). Correlation coefficient ($\rho = 0.19$) and a trendline are indicated. **B** and **C,** Two scatter plots show correlation of tumor purity with CYT and mutation burdens ($\rho = -0.70$ and -0.20, respectively). Significance was calculated using R software. **D,** Scatter plot of the correlation coefficients between gene expression and mutation burdens calculated with or without controlling for the tumor purity (*y*- and *x*-axes, respectively). Mitochondria: 175 genes, red dots; immune: 1,140 genes, blue dots (as in Fig. 1B). **E,** Purity-adjusted (yellow) and -unadjusted (green) correlations for the expression of immune marker genes (*GZMA, PRF1, PDCD1, CD274, CTLA4,* and *CD8A*) as well as *APOBEC3B* and *MLH1* with the burdens of somatic mutation. *MLH1*: the correlation was measured for three tumor types of COAD, STAD, and UCEC.

concordant with each other. To identify the molecular functions enriched with genes whose correlation with mutation abundance is subject to controlling for tumor purity, we performed GSEA, using the differential between the purity-adjusted and -unadjusted correlations as a rank metric with GO categories. Mitochondria- and immune-related molecular functions, which were identified as molecular functions enriched with genes correlated and inversely correlated with tumor purity (Table 1), were also enriched with genes whose correlation with mutation abundance became higher and lower after controlling for the tumor purity, respectively (Supplementary Table S4). Fig. 2D also shows 175 mitochondria- and 1,140 immune-related genes (from Fig. 1B) in red and blue dots, respectively, to demonstrate that mitochondria- and immune-related genes comprise major functional categories whose correlation with mutation abundance are subject to or not subject to controlling for the tumor purity. Figure 2E shows that the positive correlation coefficients with mutation burdens of six immune genes (*GZMA, PRF1, CD8A, PDCD1, CD274*, and *CTLA4*) diminished after purity adjustment. We also examined the correlation with the mutation burdens for *APOBEC3B* and *MLH1*. The activity of spontaneous cytidine deaminase, APOBEC3B, has been associated with elevated mutation rates in breast cancer and other tumor types (30), and the loss of MLH1 activity by promoter hypermethylation is associated with mutator phenotypes in some tumor types (31). As expected, a correlation and inverse correlation with the mutation abundance were observed for *APOBEC3B* and *MLH1*, respectively (for *MLH1*, correlation was measured in COAD, STAD, and UCEC tumor types; Fig. 2E). The correlation of *APOBEC3B* and inverse correlation of *MLH1* expression with mutation abundance were also affected by controlling for tumor purity but to a lesser extent compared with immune marker genes. These results suggest that using the expression of individual genes as markers for mutation abundance may be biased by tumor purity and requires caution, especially for immune-related genes.

Previous analyses reported that CYT gene expression is marginally associated with patient prognosis, in terms of overall survival (19). We further investigated the association between patient survival and CYT gene expression with or without controlling for the tumor purity. Significant association ($P < 0.05$) was observed for BRCA, LGG, LIHC, and SKCM (Supplementary Fig. S5A), but after controlling for the tumor purity, BLCA showed

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

significant association between CYT gene expression and patient survival (Supplementary Fig. S5B). Previously, a study reported that tumor purity is associated with patient survival for LGG (5). Thus, it is reasonable to speculate that the significant association between CYT gene expression and survival may be attributed to their relationship with tumor purity, at least for this tumor type. The discrepancy of the CYT–survival relationship in other tumor types may not be simply explained by tumor purity, but the tumor purity should be taken into account as a potential confounding factor in correlation analyses between clinicopathologic features, such as patient survival and expression of immune genes, such as *GMZA* and *PRF1*.

Impact of tumor purity on gene clustering

Previous investigation of the impact of tumor purity on coexpression analyses revealed that a substantial number of coexpressed gene pairs may have arisen due to their relationship with tumor purity, raising a need to account for the tumor purity in coexpression analyses (5). They also examined coexpressed gene clusters for their enriched molecular functions and the extent of correlation with the purity (5). We further examined the cluster membership of gene pairs and measured their coherence between gene clusters with or without controlling for tumor purity. For each tumor type, we first selected 5,000 highly variable genes and assigned them into 6 gene clusters by consensus clustering (purity-unadjusted). Individual gene pairs were classified into cluster-matched pairs (CM pairs) as genes in the pair belonging to the same cluster and cluster-mismatched pairs (CMM pairs) for others. The proportion of CM and CMM gene pairs was relatively constant across tumor types (CM/CMM ratios were observed in the range of 0.20–0.257; Supplementary Fig. S6). To evaluate the impact of tumor purity on gene clustering, we further performed purity-adjusted gene clustering using a partial correlation-based gene-wise distance matrix, controlling for tumor purity. According to the coherence of cluster membership, we further discriminated CM-D pairs (gene pairs whose cluster membership became discordant after purity adjustment) from CM pairs and CMM-C pairs (gene pairs whose cluster membership became concordant after purity adjustment) from CMM pairs. We calculated CM-D/CM and CMM-C/CMM ratios to evaluate the impact of tumor purity on gene clustering (Fig. 3A). High CM-D/CM ratios (>0.55) were observed for ACC, GBM, SCKM, and KICH, suggesting that more than 55% of gene pairs belonging to the same cluster would be assigned to different clusters after controlling for tumor purity. Although CESC, LICH, KIRP, and COAD tumor types were marked with low CM-D/CM ratios, a substantial number of CM gene pairs (>20%) were still found as CM-D pairs for these tumor types, suggesting that impact of tumor purity is substantial for gene clustering overall. An additional measure of CMM-C/CMM indicates the extent of how many gene pairs, which belong to different clusters, would be assigned to the same cluster after controlling for tumor purity. The CMM-C/CMM ratios were in the range of 0.05 to 0.14 and correlated with the CM-D/CM ratios (Fig. 3B), suggesting that CM-D/CM and CMM-C/CMM ratios may serve as an indicator of which tumors are more vulnerable to controlling for tumor purity in gene clustering.

To provide evidence for the need of purity adjustment, we obtained PPI gene pairs from the HPRD database (27). The frequency of PPI gene pairs was measured for four gene pair categories (CM-C, CM-D, CMM-C, and CMM-D). The PPI gene pairs were more frequent for CM-C and CMM-C compared with CM-D and CMM-D (Fig. 3C). The elevated PPI gene pair frequencies in concordant gene pairs compared with discordant pairs were largely consistent across tumor types (Supplementary Fig. S7), supporting that purity adjustment increases the functional coherence of gene clustering.

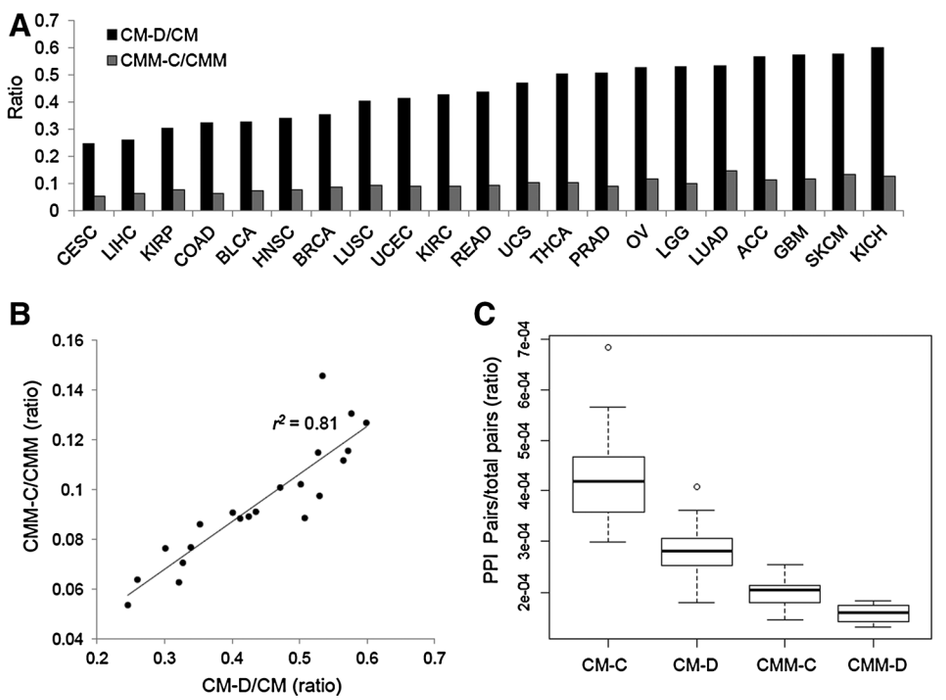

Figure 3.

The impact of tumor purity on gene clustering. **A,** CM-D/CM and CMM-C/CMM ratios are shown across 21 tumor types (black and gray bars, respectively). **B,** Scatter plot of the CM-D/CM (*x*-axis) and CMM-C/CMM (*y*-axis) ratios ($r^2 = 0.81$). **C,** Box plots of the frequency of PPI gene pairs in given gene pairs for four categories (CM-C, CM-D, CMM-C, and CMM-D) measured across 21 tumor types.

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

Impact of tumor purity on gene expression–based molecular taxonomy

It has been proposed that tumor purity may affect molecular taxonomy or tumor subtyping, based on the gene expression (5). This previous study demonstrated that the tumor purity was substantially different across known tumor subtypes (e.g., low tumor purity for GBM mesenchymal subtypes), and the correlations of the feature genes (gene classifiers) with tumor purity are distinct from overall correlations (e.g., bimodal distribution). In this study, we further investigated the extent of association between the feature genes representing individual tumor subtypes and tumor purity, as well as their potential impact on the expression-based molecular taxonomy. First, we observed that the feature genes representing four GBM subtypes were either significantly correlated (proneural and classic subtypes) or inversely correlated (neural and mesenchymal; $P < 0.001$; Kolmogorov–Smirnov tests; Fig. 4A). The unidirectional preference of GBM feature genes toward either correlation or inverse correlation may explain the previously observed bimodal distribution of the correlation between GBM feature genes and tumor purity when examined as aggregates (5). We further performed hierarchical clustering using the half of the feature genes that were either correlated or inversely correlated with tumor purity (top 400 purity-correlated and 400 purity-inversely correlated feature genes for Fig. 4B and C, respectively). The major perturbation in molecular taxonomy (the split of subtype clusters) was observed for the neural and proneural subtypes (blue and green in Fig. 4B and C, respectively). These findings suggest that the molecular taxonomy of GBM as well as the selection of feature genes may be biased by tumor purity.

We further obtained the feature genes for an additional five tumor types available for the annotations of mRNA expression-based tumor subtypes (BRCA, COAD, HNSC, LUAD, and LUSC). Feature genes representing three to four molecular subtypes per tumor type were identified and examined for their correlation with tumor purity. The extent of correlation or inverse correlation is shown as the positive and negative normalized enrichment scores, respectively (Fig. 4D). The feature genes representing mesenchymal subtypes (GBM and HNSC) and SSM (stem/serrated/mesenchymal) of COAD were substantially inversely correlated with the tumor purity, as well as the secretory subtype of LUSC. The LUSC secretory subtype was characterized by overexpression of secretory cell markers such as MUC1, as well as immune markers such as NF-κB target genes (32). Feature genes correlated with tumor purity included two classic subtypes of GBM and HNSC, as well as subtypes associated with favorable patient prognosis (proneural GBM subtype and luminal A BRCA subtype). Four sets of feature genes representing classic and mesenchymal subtypes of HNSC, SSM subtype of COAD, and secretory subtype of LUSC are selected and shown for their enrichment plots (Fig. 4E).

Discussion

In this study, we used a large-scale cancer transcriptome database from the TCGA consortium comprising over 7,794 tumor specimens and 21 tumor types. We investigated the impact of tumor purity on the expression of genes and the related issues, including the correlation between immune marker gene expression and mutation burdens, as well as gene clustering and molecular taxonomy. Because tumor purity is largely dependent on how the tumor specimen is obtained, it has been debated whether tumor purity represents biologically relevant, intrinsic tumor features or simply represents systematic biases that are determined by extrinsic features, such as surgical resection and tissue preparation. To distinguish the effects of intrinsic and extrinsic features of tumor purity, one pioneering study evaluated the correlation between hundreds of clinical features and tumor purity (5). This study failed to identify a clear association with the clinical features, and the authors concluded that tumor purity differences are largely determined by extrinsic factors and that tumor purity is a major confounding factor in cancer genome or transcriptome analyses (5). Thus, it is challenging to examine the potential impact of tumor purity on expression-based tumor analyses such as the correlative analyses of immune gene expression with clinicopathologic features, as well as expression-based gene clustering and molecular taxonomy.

Two main components of noncancerous tissues in tumor specimens are stromal and immune cells. Their expression showed comparable performance in the estimation of tumor purity in a prior study (6). Presently, the extent of inverse correlation with tumor purity was seen for the expression of immune genes. Immune cell infiltrates and the expression of immune marker genes have been highlighted as potential targets for immunotherapy. The expression of immune-related genes, a substantial fraction of which may be derived from tumor-infiltrating immune cells, has been assumed to be correlated with tumor purity. Although a number of *in silico* methods are available for the virtual microdissection of bulk tumor transcriptomes (23) and also to estimate the abundance of immune cell infiltrates (7, 8), these deconvolution-based methodologies use a gene expression matrix as an input, not dedicated to infer the origin of cells for the expression of individual genes. Technologies developed for discerning immune cell populations, such as conventional immunohistochemistry and flow cytometry combined with single-cell sequencing, can be used to investigate the abundance and cell-type–specific expression analyses (33). As demonstrated in a study (34), single-cell, sequencing-based immune profiling may alleviate the purity-associated biases. However, the issues of cost and technical concerns remain (35). Although a unique setting, such as patient-derived xenografts, enables the separation of individual transcripts from the tumor and stromal cells (22), the majority of tumor expression profiles are still obtained from bulk tumor masses, and the main purpose of our study was to highlight the potential effects of tumor purity on conventional expression-based analyses. We assumed that genes representative of cytotoxic T lymphocytes (*PRF1*, *GZMA*, and *CD8A*) and immune checkpoints (*PDCD1*, *CD274*, and *CTLA4*) may be a biased approach due to the extrinsic nature of tumor purity. Our correlative analyses revealed that immune-related genes comprise a major gene category that was inversely correlated with tumor purity, and the individual immune gene expression could be biased by tumor purity, which should be taken into account when evaluating the correlation with other clinicopathologic features including mutation burdens. A number of studies have reported the correlation between the expression of such immune-related genes and mutation burdens (7, 15, 19, 36). However, our results suggest that such correlations may also be biased by tumor purity. In this view, the correlation between immune marker gene expression and mutation burdens may be due to their correlation with tumor purity. Genes whose expression is inversely correlated with tumor purity will also correlate with mutation burdens to some extent,

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

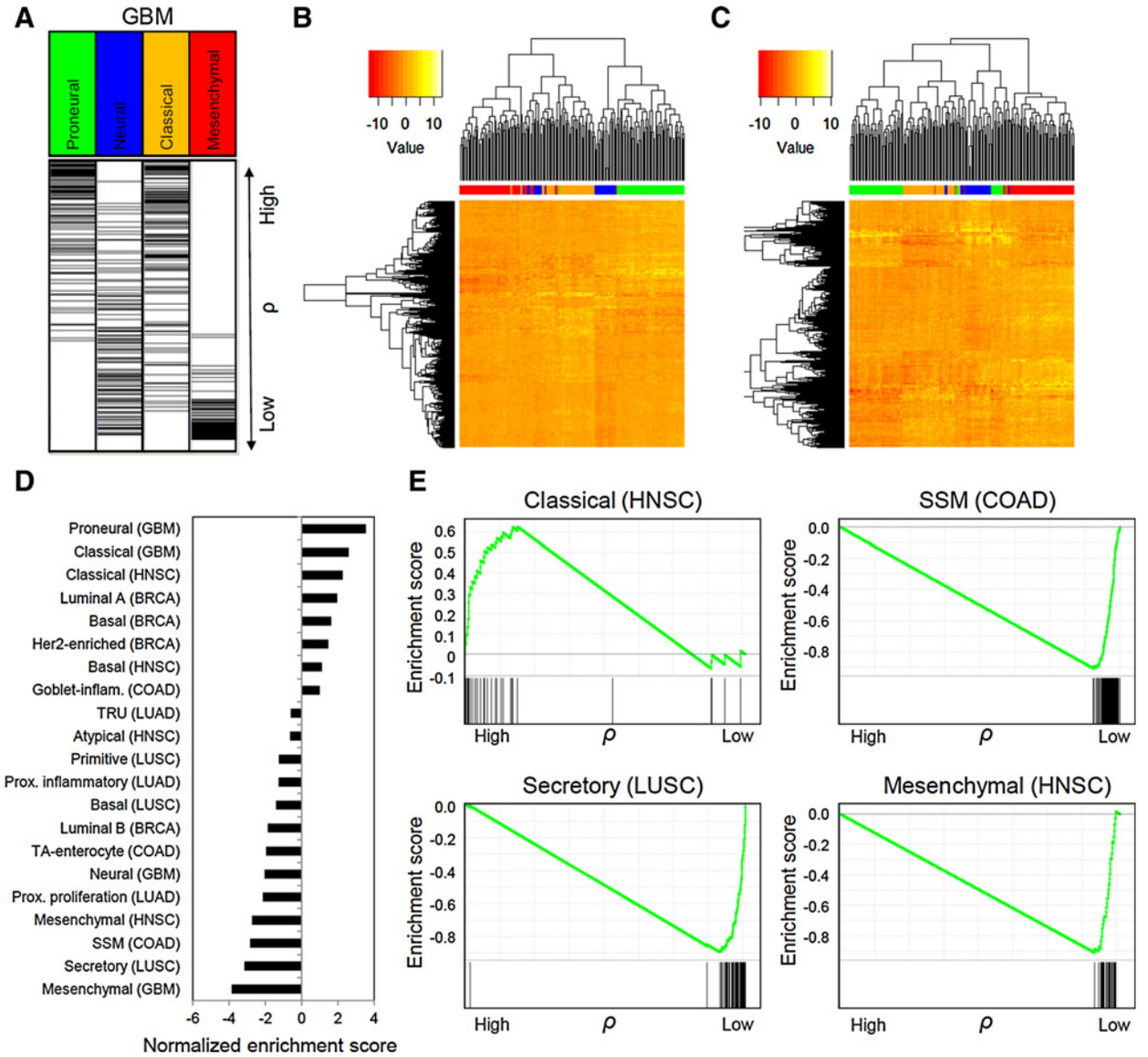

Figure 4.

The impact of tumor purity on the molecular taxonomy. **A,** Correlation of feature genes for four GBM subtypes with tumor purity. Enrichment towards high or low correlation is indicated (right). $P < 0.001$ by Kolmogorov–Smirnov test. **B** and **C,** Hierarchical clustering of a subset of 400 feature genes correlated with tumor purity (**B**), and 400 feature genes inversely correlated with tumor purity (**C**). Color bars above the heatmap represent four GBM subtypes (green: proneural, blue: neural, orange: classic, and red: mesenchymal). **D,** For six tumor types (GBM, HNSC, BRCA, COAD, LUAD, and LUSC), the feature genes representing three to four tumor subtypes of given tumor type were selected. Normalized enrichment scores, as the extent of enrichment toward correlation and inverse correlation with tumor purity, are shown. **E,** Four examples of classic (HNSC), SSM (COAD), secretory (LUSC), and mesenchymal (HNSC) subtypes are shown for their enrichment of feature genes toward either correlated (left; high ρ) and inversely correlated (right; low ρ) with tumor purity as snapshots of enrichment plots of GSEA output. Green line: running sum statistic to estimate enrichment score.

regardless of their biological functions. Immune gene expression along with mutational burdens may be interdependent with each other (15). However, it is expected that some of the observed dependence between parameters may have arisen due to tumor purity. The need to adjust tumor purity when evaluating the correlation between immune gene expression or the abundance of immune cell infiltrates with the mutation abundance has been previously proposed (7, 36). The current findings provide solid evidence by evaluating the correlation with or without control for tumor purity.

A pioneering study on tumor purity previously described the potential confounding effects of tumor purity in terms of coexpression network, molecular taxonomy, and differential expression between tumor and nontumor tissues (5). In this study, we

Published OnlineFirst November 15, 2017; DOI: 10.1158/2326-6066.CIR-17-0201

examined the coherence of cluster membership for individual gene pairs with or without controlling for tumor purity and also investigated the correlation with tumor purity for feature genes representing tumor subtypes. First, we observed that gene clustering controlled for tumor purity can alter the cluster membership for a substantial number of gene pairs. Specifically, ratios of CM-D/CM and CMM-C/CMM gene pairs were observed in the rage of 0.24 to 0.59 and 0.05 to 0.15, respectively. These two ratios correlated with each other and varied across tumor types. For example, in LUAD with high CM-D/CM and CMM-C/CMM ratios, 53% of gene pairs that belong to the same cluster lost the cluster membership concordance, whereas 14% of gene pairs were assigned to the same cluster after the purity adjustment. Even for the tumor type with the lowest ratios such as CESC (CM-D/CM and CMM-C/CMM ratios as 0.24 and 0.05, respectively), purity adjustment substantially altered the cluster membership in gene clustering. We also demonstrated that concordant gene pairs (genes in pairs belonging to the same cluster after purity adjustment) were more frequent in the number of PPI gene pairs, suggesting that purity adjustment may improve the functional coherence of gene clusters.

Next, we investigated the extent of correlation with tumor purity for feature genes used for molecular taxonomy. We demonstrated that feature genes representing the tumor subtypes were often either correlated or inversely correlated with tumor purity. As expected, inverse correlation with tumor purity was observed for feature genes representing mesenchymal subtypes (COAD, GBM, and HNSC). This is expected because stromal genes constitute the major fraction of feature genes representing mesenchymal subtypes. Hierarchical clustering using a subset of feature genes, either correlated or inversely correlated with the tumor purity, showed certain subtypes (classic and mesenchymal GBM subtypes) were relatively robust compared with other subtypes (proneural and neural GBM subtypes), suggesting that the impact of tumor purity may be context dependent. The feature genes representing classic subtypes were correlated with tumor purity for GBM and HNSC. We assumed that classic subtypes were characterized by recurrent alterations of the corresponding tumor types, such as 7p/q amplifications (GBM) and 3q amplifications (HNSC), and that this representative nature of classic subtypes may be responsible for the positive correlation of feature genes with tumor purity. However, further investigation of why some tumor subtypes with favorable prognosis (proneural/GBM and luminal A/BRCA) shows positive correlation with tumor purity is needed.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: T.-M. Kim
Development of methodology: J.-K. Rhee, Y.C. Jung, T.-M. Kim
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.-K. Rhee, Y.C. Jung, K.R. Kim, J. Yoo, J. Kim, Y.-J. Lee, B.C. Cho
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.-K. Rhee, Y.C. Jung, K.R. Kim, J. Yoo, J. Kim, Y.-J. Lee, H.H. Lee, T.-M. Kim
Writing, review, and/or revision of the manuscript: Y.H. Ko, B.C. Cho, T.-M. Kim
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y.H. Ko, H.H. Lee, B.C. Cho
Study supervision: Y.H. Ko, B.C. Cho, T.-M. Kim

Acknowledgments

This work was supported by the Korea Health Technology R&D Project via the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant nos. HI15C1578, HI15C1592, and HI15C3224).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 18, 2017; revised August 30, 2017; accepted November 8, 2017; published OnlineFirst November 15, 2017.

References

1. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nat Rev Cancer 2009;9:239–52.
2. Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature 2013;501:346–54.
3. Pages F, Galon J, Dieu-Nosjean MC, Tartour E, Sautes-Fridman C, Fridman WH. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene 2010;29:1093–102.
4. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science 2011;331:1565–70.
5. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971.
6. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013;4:2612.
7. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 2016;17:174.
8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7.
9. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med 2014;371:2189–99.
10. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348:124–8.
11. Robert C, Long GV, Brady B, Dutriaux C, Maio M, Mortier L, et al. Nivolumab in previously untreated melanoma without BRAF mutation. N Engl J Med 2015;372:320–30.
12. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627–39.
13. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 2015;372:2509–20.
14. Robbins PF, Lu YC, El-Gamil M, Li YF, Gross C, Gartner J, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat Med 2013;19:747–52.
15. Danilova L, Wang H, Sunshine J, Kaunitz GJ, Cottrell TR, Xu H, et al. Association of PD-1/PD-L axis expression with cytolytic activity, mutational load, and prognosis in melanoma and other solid tumors. Proc Natl Acad Sci U S A 2016;113:E7769–E77.
16. Herbst RS, Soria JC, Kowanetz M, Fine GD, Hamid O, Gordon MS, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature 2014;515:563–7.
17. Taube JM, Klein A, Brahmer JR, Xu H, Pan X, Kim JH, et al. Association of PD-1, PD-1 ligands, and other features of the tumor immune

microenvironment with response to anti-PD-1 therapy. Clin Cancer Res 2014;20:5064–74.
18. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 2015;350:207–11.
19. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;160:48–61.
20. Ock CY, Keam B, Kim S, Lee JS, Kim M, Kim TM, et al. Pan-cancer immunogenomic perspective on the tumor microenvironment based on PD-L1 and CD8 T-cell infiltration. Clin Cancer Res 2016;22:2261–70.
21. Elloumi F, Hu Z, Li Y, Parker JS, Gulley ML, Amos KD, et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med Genom 2011;4:54.
22. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, et al. Stromal contribution to the colorectal cancer transcriptome. Nat Genet 2015;47:312–9.
23. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SG, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet 2015;47:1168–78.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
25. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
26. Kim S. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods 2015;22:665–74.
27. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database–2009 update. Nucleic Acids Res 2009;37:D767–72.
28. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
29. Dabney AR. ClaNC: point-and-click software for classifying microarrays to nearest centroids. Bioinformatics 2006;22:122–3.
30. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 2013;494:366–70.
31. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 2013;155:858–68.
32. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res 2010;16:4864–75.
33. Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. Nat Rev Genet 2016;17:441–58.
34. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017;8:15081.
35. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res 2016;5.
36. Varn FS, Wang Y, Mullins DW, Fiering S, Cheng C. Systematic pan-cancer analysis reveals immune cell interactions in the tumor microenvironment. Cancer Res 2017;77:1271–82.

AACR American Association for Cancer Research

Cancer Immunology Research

Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types

Je-Keun Rhee, Yu Chae Jung, Kyu Ryung Kim, et al.

Cancer Immunol Res 2018;6:87-97. Published OnlineFirst November 15, 2017.